

NOISE-ROBUST ASR FOR THE THIRD ‘CHiME’ CHALLENGE EXPLOITING TIME-FREQUENCY MASKING BASED MULTI-CHANNEL SPEECH ENHANCEMENT AND RECURRENT NEURAL NETWORK

Zaihu Pang, Fengyun Zhu

Lingban Technology Co., Ltd.
Beijing, China
{zhpang, fyzhu}@ling-ban.com

ABSTRACT

In this paper, the Lingban entry to the third ‘CHiME’ speech separation and recognition challenge is presented. A time-frequency masking based speech enhancement front-end is proposed to suppress the environmental noise utilizing multi-channel coherence and spatial cues. The state-of-the-art speech recognition techniques, namely recurrent neural network based acoustic and language modeling, state space minimum Bayes risk based discriminative acoustic modeling, and i-vector based acoustic condition modeling, are carefully integrated into the speech recognition back-end. To further improve the system performance by fully exploiting the advantages of different technologies, the final recognition results are obtained by lattice combination and rescoring. Evaluations carried out on the official dataset prove the effectiveness of the proposed systems. Comparing with the best baseline result, the proposed system obtains consistent improvements with over 57% relative word error rate reduction on the real-data test set.

Index Terms— Noise-robust ASR, Multi-channel speech enhancement, Time-frequency masking, Recurrent neural network, CHiME challenge

1. INTRODUCTION

Far-field noise-robust automatic speech recognition (ASR) in real-world environments is still a challenging problem. The series of ‘CHiME’ speech separation and recognition challenges offered a great opportunity for the researchers from the signal processing community and the ASR community to work collaboratively toward this goal. The past ‘CHiME’ challenges have contributed to the development of several speech enhancement techniques, and novel framework that integrate speech enhancement and ASR [1, 2].

The main difference between the current ‘CHiME’ challenge and the past ones is instead of working on simulated data only, the current challenge focuses on real-world data: multi-channel recording using mobile tablet device in a variety of noisy public environments [3]. How to make use of

both the real and simulated data to improve the system performance remains an open question for the current challenge.

In this paper, we presented the Lingban entry to the 3rd ‘CHiME’ speech separation and recognition challenge. A time-frequency masking based speech enhancement front-end is proposed to suppress the environmental noise utilizing multi-channel coherence and spatial cues. An adaptive microphone array self-calibration method is adopted to overcome the problem of microphone mismatch. For acoustic modeling, neural network acoustic models including maxout neural network [4] and long short term memory with project layers (LSTMP) [5, 6] are adopted. Mel-frequency cepstral coefficient (MFCC) based feature-space maximum likelihood linear regression (fMLLR) features are utilized. Moreover, online extracted i-vector is also used as the network input for encoding these effects: speaker, channel and background noise [7, 8, 9, 10]. State-space Minimum Bayes Risk (sMBR) discriminative training is conducted on the neural networks based acoustic models [11, 12]. For language and lexicon modeling, to model the inter-word silence more precisely, pronunciation lexicon with silence probability is adopted [13]. 4-gram language model with Kneser-Ney smoothing is adopted in the first-pass decoding. A language model based on jointly trained recurrent neural network and maximum entropy models (RNNME) is adopted for the second-pass rescoring [14, 15]. Finally, to further improve the system performance by fully exploiting the advantages of different technologies, the recognition results are obtained by lattice combination and rescoring.

Evaluations are carried out on the official dataset. Comparing with the best baseline result, the proposed system obtains consistent improvements with over 57% and 42% relative word error rate (WER) reduction on real and simulated test set respectively.

The rest of this paper is organized as follows. The time-frequency masking based speech enhancement front-end is presented in Section 2. The speech recognition back-end is presented in Section 3. Experiments are presented in Section 4, followed by the conclusions in Section 5.

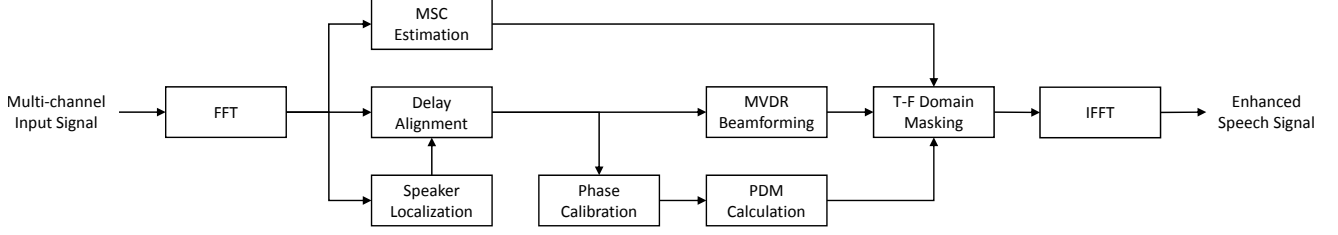


Fig. 1. Schematic diagram of the proposed speech enhancement method.

2. SPEECH ENHANCEMENT

Fig.1 shows a schematic diagram of the proposed speech enhancement method. The proposed method is developed on the basis of the speech enhancement baseline of the ‘CHiME’ challenge. Keeping the analysis-synthesis scheme, the microphone failure detector, the SPR-PHAT based speaker localizer and the MVDR beamformer identical to the baseline, in this study, time-frequency masking based on multi-channel magnitude-squared coherence (MSC) and phase difference measurement (PDM) is introduced. The MSC-based masking is used to suppress the diffused noise, while the PDM-based masking is used to suppress the directional interference not coming from the target direction. The time-frequency masking is performed by filtering the output subband signal of the MVDR beamformer with the said maskers. Since the phase difference based methods are sensitive to the mismatch of microphone phase response, an adaptive microphone array self-calibration method is adopted.

2.1. MSC based time-frequency masking

The MSC between two signals $x(t)$ and $y(t)$ is defined as:

$$C_{xy}(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)}, \quad (1)$$

where f is the frequency, $S_{xy}(f)$ is the cross-spectral density between the two signals, $S_{xx}(f)$ and $S_{yy}(f)$ are the auto-spectral density of x and y respectively. The values of the coherence function, which always satisfy $0 \leq C_{xy}(f) \leq 1$, indicates the extent to which the power of y could be predicted from x by a linear system. In multi-channel signal processing, the MSC is an efficient mean of noise reduction [16]. In the case of microphone array with sufficiently large microphone spacing, incoherent (diffused) noise would be indicated by small coherence values, while the directional signals would have large coherence values [17, 18].

In this study, MSC between all the microphone pairs (except the channels detected to be failed) are estimated by Welch’s method of periodogram. The MSC-based time-frequency masker is derived by averaging the MSC over all the microphone pairs.

2.2. PDM based time-frequency masking

In the past ‘CHiME’ challenges, phase difference based time-frequency masking noise suppression front-ends was proved to be effective in the case of dual-channel simulated data [19]. After steering the multi-channel signals towards the target direction by delay alignment, inter-channel phase differences could be obtained. Time-frequency bins which has a phase difference not close to zero are less likely to be the ones from the target direction.

In the current ‘CHiME’ challenge, the experimental condition is extended from dual-channel simulated data to 6-channel real-world recordings. For the multi-channel sub-band input signals, delay alignment is performed by steering the signals towards the target direction given by the speaker localizer. Phase differences between all the microphone pairs (except the 2nd channel and the failed ones) could be calculated for all the frequency bins. The PDM could be obtained by averaging the absolute value of the phase differences over all the microphone pairs. Assuming that the estimated speaker location is correct, the bins dominated by the signal coming from the target direction would have small PDM values.

Since the PDM values don’t lies within the range of $[0, 1]$, the PDM based time-frequency masker $W'_P(f, t)$ is obtained by performing a non-linear transformation on the PDM value $W_P(f, t)$, and hard clipped to have a maximum value of 1:

$$W'_P(f, t) = \text{clip}(1 - \tanh(W_P(f, t) - \alpha(f))), \quad (2)$$

where $\alpha(f)$ is a bias function, which is introduced to fit the frequency-dependent distribution of the PDM values. $\alpha(f)$ is empirically determined by dividing the PDM values in the training set into signal and noise, prior to the testing stage. In this study, $\alpha(f) = 0.4 + 0.3f/f_s$, where f_s is the sampling rate, is used.

2.3. Microphone array self-calibration

It’s well known that microphone mismatch could cause speech signal cancellation in adaptive beamforming [20]. Similarly, in the context of phase difference based time-frequency masking, the accuracy of the masking function would be degraded by the interference of the unknown phase response of the microphones, and the environment-dependent

phase response of the transfer function between the speaker and the microphones.

In this study, adaptive microphone array self-calibration with recursive configuration is adopted [21]. For each channel, a transfer function is estimated by minimizing the error between the output signal of a delay-and-sum beamformer and the delay aligned microphone signal. To evaluate the effectiveness of the self-calibration in the context of phase difference base time-frequency masking, only the phase response of the transfer function is used, leaving the MVDR beamformer identical to the baseline.

In this study, a 2-stage self-calibration scheme is proposed. In the first stage, an off-line calibration is done on the training set, which provides a robust estimation of the microphone phase response, due to the relatively large amount of training data. The parameters of the calibration filters are updated in a batched fashion, using the frames with SNR greater than the utterance-level median SNR. The SNR is estimated using the method provided by the acoustic simulation baseline. In the second stage, after the phase response of the first calibration filter is compensated, an on-line calibration is done for each utterance in the test set, which provides an estimation of the environment-dependent transfer function. The parameters of the calibration filters are updated in a batched fashion, using all the frames in the evaluation utterance. The phase responses of the two calibration filters are compensated before the PDM calculation.

3. SPEECH RECOGNITION

In recent years, the recurrent neural networks (RNNs) based speech recognition systems have brought about noteworthy performance improvements. Specifically for acoustic modeling, the long short-term memory (LSTM) based deep networks have been shown to give the state-of-the-art performance on some of the speech recognition tasks. In the seminal work, Graves *et al.* [22] proposed to use stacked bidirectional LSTM trained with connectionist temporal classification [23] for phoneme recognition. Subsequently, LSTM RNNs have been successfully applied and shown to give state-of-the-art performance on robust speech recognition task [24], and many large vocabulary speech recognition tasks [5, 6]. On the other hand, RNN based language models have also shown advantages over the conventional N-gram language models in both perplexity and speech recognition error rate.

The speaker variability is also an important issue. Efforts have been made to train acoustic models using speaker-adapted features, which can be obtained by speaker normalization techniques such as the vocal tract length normalization (VTLN) and fMLLR [25]. Another group of approaches are using speaker-aware training [26], where the speaker information, such as speaker code [27] and speaker- or utterance-level i-vector [9, 10], is provided to the neural networks directly. It should be noticed that the information about channel and

background noise is also encoded by the i-vector.

In this study, Mel-scale log-filterbank coefficients (FBANK) features, MFCC-based fMLLR features are used. The i-vector was extracted online and concatenated to the acoustic feature. The hybrid approach for acoustic modeling is adopted, in which the neural networks outputs are converted to pseudo likelihood, and used as the state output probability of the HMMs. The networks are trained based on alignments from GMM-HMM systems. Neural network acoustic models including maxout neural network [4] and LSTMP [5, 6] are adopted. sMBR discriminative training is conducted on the neural networks based acoustic models [11, 12]. To model the inter-word silence more precisely, pronunciation lexicon with silence probability is adopted [13]. The baseline 3-gram language model is replaced by a 4-gram language model with Kneser-Ney smoothing trained on the official training set. An RNNME based language model [14, 15] is adopted for the second-pass rescoring. Finally, system combination is introduced to further improve the system performance by integrating systems with different properties using lattice combination and rescoring.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Experimental setups

All the experiments were conducted on the official dataset. The GMM-HMM system was trained using Kaldi [28]. All the networks were trained based on the alignment result of the GMM-HMM system with 2824 tied context dependent HMM states. In the training procedure of LSTM networks, the strategy introduced by [29] was applied to scale down the gradients. Besides, since the information from the future frames helps the LSTM networks making better decisions, we also delayed the output HMM state labels by 4 frames. The feed-forward DNNs used the concatenated features, which were produced by concatenating the current frame with 7 frames in its left and right context. The inputs to LSTMP networks were only the current frames.

4.2. Speech enhancement

Table 1 shows the evaluation results of the proposed speech enhancement methods with the baseline acoustic model (GMM-HMM) and language model (3-gram). Comparing with the baseline using noisy data, the performance of the baseline with speech enhancement front-end is greatly improved on simulated test set, but degraded on real test set. This phenomenon shows that mismatch between enhanced real and simulated data is introduced by the baseline beamforming based speech enhancement process. The system performance on the real test set would be greatly influenced by this mismatch, since relatively large amount of simulated data is used in the training phase.

Table 1. Evaluation results (WER[%]) of the proposed speech enhancement front-ends with the baseline acoustic model (GMM-HMM) and language model (3-gram). “CAL” means phase calibration.

Speech Enh.	Dev. set		Test set	
	Real	Simu.	Real	Simu.
Baseline (noisy)	18.70	18.71	33.23	21.59
Baseline (enhanced)	20.55	9.79	37.36	10.59
MSC	15.12	9.57	26.42	12.01
PDM	13.26	8.23	26.23	10.20
MSC+PDM	12.45	8.45	24.72	11.27
PDM+CAL	13.07	7.96	25.11	10.33
MSC+PDM+CAL	12.14	8.71	24.55	11.51

By introducing the MSC based time-frequency masking into the baseline speech enhancement front-end, the WER on real test set is reduced by 10.94%. It proves the effectiveness of the MSC based time-frequency masking. By contrast, the WER on simulated test set is increased by 1.42%. It shows that this method performs more equally on real and simulated data. Hence the said mismatch is weakened. The effectiveness of the PDM based time-frequency masking is also proved by the WER reduction on real test set of 11.13%. While the WER on simulated test set is lower than the MSC based system. Comparing to the MSC based method, although better performance on real test set is achieved, greater mismatch between real and simulated data is introduced by the PDM based method. By using the two kinds of time-frequency masking together, an absolute WER reduction of 12.64% is achieved.

It’s also shown that the by introducing phase calibration, further absolute WER reduction of 1.12% and 0.17% is made on the basis of the “PDM” and “MSC+PDM” systems respectively. The mismatch between real and simulated data is further weakened. It should be noticed that, since the time for the challenge is limited, further experiments for the phase calibration is *not* conducted. In the following experiments, the “MSC+PDM” front-end is used.

4.3. Lexicon and language modeling

Table 2 shows the evaluation results of different language models with the proposed front-end (MSC+PDM) and the baseline acoustic model (GMM-HMM). The N-gram language models are used in the first-pass decoding, while the RNNME-based language model is used for the second-pass rescoring.

Experiment results show that the system performance is improved consistently by introducing pronunciation lexicon with silence probability. The system performance is further improved by introducing the 4-gram language model and RNNME-based language model.

Table 2. Evaluation results (WER[%]) of different language models with the proposed front-end (MSC+PDM) and the baseline acoustic model (GMM-HMM). “SP” means pronunciation lexicon with silence probability. “RNNME” means RNNME-based language model.

Language model	Dev. set		Test set	
	Real	Simu.	Real	Simu.
Baseline (3-gram)	12.45	8.45	24.72	11.27
3-gram+SP	11.70	8.23	23.79	11.11
4-gram+SP	9.93	7.59	20.42	9.96
4-gram+SP+RNNME	9.38	7.05	19.68	9.25

Table 3. Evaluation results (WER[%]) of different features and acoustic models with the proposed front-end (MSC+PDM) and language model (4-gram+SP). “dp” means dropout. “fM” and “iV” in the lower part of the table means fMLLR and i-vector for short.

Feature	Model	Dev. set		Test set	
		Real	Simu.	Real	Simu.
FBank	Maxout	10.96	7.75	21.33	9.68
FBank	LSTMP	10.96	7.02	20.82	9.21
fMLLR	Maxout	9.44	7.08	18.41	8.57
fMLLR	Maxout+dp	9.37	6.86	18.55	8.59
fMLLR	LSTMP	8.88	6.07	16.86	7.56
fM+iV	Maxout+dp	9.37	6.89	18.42	8.61
fM+iV	LSTMP	8.32	5.86	16.45	7.20
fM+iV	Maxout+dp+sMBR	8.71	6.42	17.60	8.09
fM+iV	LSTMP+sMBR	7.91	5.43	15.55	6.86

4.4. Acoustic modeling

Table 3 shows the evaluation results of different features and different acoustic models with the proposed front-end (MSC+PDM) and language model (4-gram+SP). It should be noted that the RNNME based second-pass rescoring is not used in this evaluation.

The maxout deep neural network has 4 hidden layers and each layer has 1000 neurons with a group size of 4. The LSTMP recurrent neural network has a single hidden layer with 1000 neurons and 700 projection units. The 40-dimensional fMLLR features are attained based on the 13-dimensional MFCC features. The fMLLR transformation matrix is estimated using the baseline GMM-HMM. The 50-dimensional i-vector features are extracted on utterance level. The language model used in the sMBR discriminative training is the baseline 3-gram language model.

Experimental result shows that the LSTMP based method performs better than the maxout based method while using the same features. While using same acoustic modeling methods, fMLLR based feature performs better than Fbank based feature. By combining the fMLLR based feature with

Table 4. Evaluation results (WER[%]) of the combined systems.

System	Dev. set		Test set	
	Real	Simu.	Real	Simu.
Maxout+dp+sMBR, 4-gram+SP, RNNME	8.06	5.90	16.93	7.59
LSTMP+sMBR, 4-gram+SP, RNNME	7.20	4.85	14.69	6.27
Maxout+dp+sMBR, LSTMP+sMBR, 4-gram+SP, RNNME	7.06	4.86	14.28	6.14

the i-vector based feature, the system performance improved consistently. By using sMBR based discriminative training, the system performance improved consistently. The best WER of 15.55% is achieved by the “LSTMP+sMBR” on real data, and the second best acoustic modeling method is “Maxout+dp+sMBR”.

Table 4 shows the evaluation results of the combined systems. Since the RNNME based second-pass rescoring is not included in the evaluation of acoustic modeling, the RNNME based second-pass rescoring is introduced into the two best-performed acoustic models (Maxout+dp+sMBR, LSTMP+sMBR) with the 4-gram+SP language model in one-pass decoding. Moreover, the lattice combination is performed on the two intergraded systems to achieve a better performance. The front-end used in this evaluation is MSC+PDM. The feature used in this evaluation is fMLLR+i-vector.

Experimental result shows that the “LSTMP+sMBR” and “Maxout+dp+sMBR” systems obtain further improvement by introducing RNNME-based language model second-pass rescoring. Through lattice combination, a final WER of 14.28% is achieved on the real test set.

4.5. Overall comparison

Table 5 shows the results of the overall comparison of the proposed systems and the baseline systems:

- Baseline I: the GMM-HMM baseline on noisy data;
- Baseline II: the GMM-HMM baseline on enhanced data;
- System I: proposed speech enhancement front-end with baseline acoustic model (GMM-HMM) and language model (3-gram);
- System II: proposed language model with baseline acoustic model (GMM-HMM) on noisy data;
- System III: proposed acoustic model with baseline language model (3-gram) on noisy data;
- System IV: the best-performed proposed system.

Table 6 shows the WERs for the 4 acoustic environments achieved by the best system.

Table 5. Overall comparison of the evaluation results (WER[%]) of the proposed systems and the baseline systems.

System	Development set		Test set	
	Real	Simu.	Real	Simu.
Baseline I	18.70	18.71	33.23	21.59
Baseline II	20.55	9.79	37.36	10.59
System I	12.14	8.71	24.55	11.51
System II	15.55	15.49	28.02	18.26
System III	12.38	10.91	21.47	14.09
System IV	7.06	4.86	14.28	6.14

Table 6. WERs[%] for the 4 acoustic environments achieved by the best system.

Environment	Development set		Test set	
	Real	Simu.	Real	Simu.
BUS	8.87	4.01	16.19	4.45
CAF	5.50	6.05	13.37	6.48
PED	5.69	4.25	17.02	6.05
STR	8.17	5.12	10.53	7.58

Experimental results from the upper part of table 5 show the contribution of the proposed front-end, acoustic model and language model separately. Comparing to the baseline system on enhanced data, by introducing the proposed speech enhancement method, an absolute WER reduction of 12.81% is achieved. Comparing to the baseline systems using noisy data, absolute WER reduction of 11.76% and 5.21% is achieved by the proposed acoustic model and language model respectively. The system IV, which make use of the proposed front-end and back-end, performs the best result. A final WER of 14.28% is achieved on the real test set.

5. CONCLUSIONS

In this paper, we presented the Lingban entry to the 3rd ‘CHiME’ speech separation and recognition challenge. A time-frequency masking based speech enhancement front-end is proposed. The state-of-the-art recurrent neural networks based acoustic and language modeling methods are adopted. Evaluations are carried out on the official dataset. Comparing with the best baseline result, the proposed system obtains consistent improvements with over 57% relative WER reduction.

Since the front-end and back-end are working separately in the current implementation, a unified end-to-end learning framework for multi-channel noise-robust ASR would be proposed in the future works. Furthermore, the proposed methods should be evaluated on real-application tasks, such as spontaneous speech recognition using general purpose commercial mobile devices with less microphones.

6. ACKNOWLEDGMENT

The authors would like to thank Zhiping Zhang, Xiangang Li, Yi Liu and Tong Fu for their kindly helps.

7. REFERENCES

- [1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, May 2013.
- [2] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, “The second CHiME speech separation and recognition challenge an overview of challenge systems and outcomes,” in *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.
- [3] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.
- [4] P. Swietojanski, J. Li, and J. Huang, “Investigation of maxout networks for speech recognition,” in *Proceedings of the 2014 ICASSP*, 2014, pp. 7649–7653.
- [5] H. Sak, A. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” in *Proceedings of the 2014 INTERSPEECH*, 2014, pp. 338–342.
- [6] X. Li and X. Wu, “Constructing long short-term memory based deep recurrent neural network for large vocabulary speech recognition,” in *Proceedings of the 2015 ICASSP*, 2015.
- [7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [8] O. Glembek, L. Burget, P. Matejka, M. Karafiát, and P. Kenny, “Simplification and optimization of i-vector extraction,” in *Proceedings of the 2011 ICASSP*, 2011, pp. 4516–4519.
- [9] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” in *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 55–59.
- [10] A. Senior and I. Lopez-Moreno, “Improving dnn speaker independence with i-vector inputs,” in *Proceedings of the 2014 ICASSP*, 2014, pp. 225–229.
- [11] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” in *Proceedings of the 2013 INTERSPEECH*, 2013.
- [12] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, “Sequence discriminative distributed training of long short-term memory recurrent neural networks,” in *Proceedings of the 2014 INTERSPEECH*, 2014, pp. 1209–1213.
- [13] G. Chen, H. Xu, M. Wu, D. Povey, and S. Khudanpur, “Pronunciation and silence probability modeling for ASR,” in *Proceedings of the 2015 INTERSPEECH*, 2015.
- [14] T. Mikolov, *Statistical language models based on neural networks*, Ph.D. thesis, Brno University of Technology, 2012.
- [15] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Černocký, “RNNLM - recurrent neural network language modeling toolkit,” in *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 196–201.
- [16] R. Le Bouquin and G. Faucon, “Using the coherence function for noise reduction,” *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, no. 3, pp. 276–280, June 1992.
- [17] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer-Verlag, New York, NY, 2001.
- [18] M. B. Trawicki and M. T. Johnson, “Distributed multichannel speech enhancement with minimum mean-square error short-time spectral amplitude, log-spectral amplitude, and spectral phase estimation,” *Signal Processing*, vol. 92, no. 2, pp. 345–356, Feb. 2012.
- [19] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, “Discriminative methods for noise robust speech recognition: A CHiME Challenge Benchmark,” in *The 2nd CHiME Workshop on Machine Listening in Multisource Environments*, Vancouver, Canada, June 2013, pp. 19–24.
- [20] S. Haykin and K. J. R. Liu, *Handbook on Array Processing and Sensor Networks*, Wiley-IEEE Press, Jan. 2010.
- [21] M. Buck, T. Haulick, and H. Pfeleiderer, “Self-calibrating microphone arrays for speech signal acquisition: A systematic approach,” *Signal Processing*, vol. 86, no. 6, pp. 1230–1238, June 2006.
- [22] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proceedings of the 2013 ICASSP*, 2013, pp. 6645–6649.

- [23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.
- [24] J. Geiger, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, “Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling,” in *Proceedings of the 2014 INTERSPEECH*, 2014, pp. 631–635.
- [25] F. Seide, G. Li, X. Chen, and D. Yu, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” in *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 24–29.
- [26] D. Yu and L. Deng, *Automatic speech recognition - A deep learning approach*, Springer-Verlag London, 2015.
- [27] O. Abdel-Hamid and H. Jiang, “Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code,” in *Proceedings of the 2013 ICASSP*, 2013, pp. 7942–7946.
- [28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, “The Kaldi speech recognition toolkit,” in *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.
- [29] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” *CoRR*, vol. 1211.5063, 2012.